知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

甲骨文识别技术研究现状与展望

刘洋[1,2] 陆逸[1] 魏钰驰[1] 孙智莹[1] 朱立芳[3]
[1] 武汉大学信息管理学院 武汉 430072
[2] 武汉大学大数据研究院 武汉 430072
[3] 广东财经大学人文与传播学院 广州 510320

摘要：[**目的 / 意义**] 对数字人文视域下甲骨文识别研究进行系统性综述，为后续研究提供参考和借鉴，推动数字人文研究有效发展与古籍文字识别利用。[**方法 / 过程**] 采用文献计量分析的方法，在 WOS、中国知网等多个学术平台检索文献，共筛选 103 篇英文文献和 52 篇中文文献进行综述。[**结果 / 结论**] 从传统识别技术、机器学习和深度学习 3 个层面解读甲骨文识别研究现状，但并未深入阐述识别算法机制。甲骨文识别技术由传统的特征提取逐渐转为基于深度学习的识别技术，在识别精度等方面有很大提升，但仍存在一些不足，同时甲骨文知识库、知识图谱的构建与领域知识的建立在该领域有较好的发展潜力。

关键词：数字人文 甲骨文识别 研究进展 系统性综述

分类号：G203

引用格式：刘洋 , 陆逸 , 魏钰驰 , 等 . 甲骨文识别技术研究现状与展望 [J/OL]. 知识管理论坛 , 2022, 8(2): 115-125[引用日期]. http://www.kmf.ac.cn/p/337/.

伴随着数字技术与人文研究碰撞的不断深入，作为交叉领域的“数字人文”研究其地位日益凸显。数字人文借助信息技术、数字技术助力传统人文学科研究，成为当下“新文科”发展的新生长点 [1-2]。数字人文研究涉及多个领域，研究对象为人文学科领域各类可数字化的资源 [3]，形式上包括图像资料、无格式文本、视频音频等，内容上包括历史文献、图书档案等。数字人文研究在文学、语言学、历史、地理等多个领域发挥重要作用。

古籍数字化是数字人文研究最基础的条件之一 [4]，甲骨文识别研究作为古籍数字化的重要一环，也是数字人文的研究对象，在古籍特定领域数字人文研究中具有重要意义。在“数字人文”理念和技术的帮助下，甲骨文等古籍文字资源的挖掘整合、特征提取、识别研究等都能在深度与广度上得到拓展，帮助古籍文字资源成为兼具历史性、可视性且组织结构合理

基金项目：本文系国家自然科学基金青年项目“突发公共卫生事件公众心理应激信息表征及干预机制研究”（项目编号：72204190）、教育部人文社科项目青年项目“基于社交机器人的突发公共卫生事件公众心理应激干预研究”（项目编号：22YJCZH114）和中国博士后面上基金“突发公共卫生事件公众心理应激信息表征及干预机制研究”（项目编号：2022M722476）研究成果之一。

作者简介：刘洋，助理教授，博士，E -mail: yang.liu27@whu.edu.cn；陆逸，本科生；魏钰驰，本科生；孙智莹，本科生；朱立芳，讲师，博士。

收稿日期：2023-01-28 发表日期：2023-04-21 本文责任编辑：刘远颖

的数字人文记忆。

甲骨文是迄今为止发现的最早具有完整体系的汉字 [5]，具有深厚的历史文化意义。2017 年甲骨文入选联合国教科文组织“世界记忆名录”，其重要的文化价值和历史意义得到世界认可。习总书记在 2019 年为纪念甲骨文发现 120 周年座谈会所发贺信中提及“殷墟甲骨文的重大发现在中华文明乃至人类文明发展史上具有划时代的意义”，强调要确保甲骨文研究有人做、有传承。综合运用人工智能等技术手段进行甲骨文识别，促进其在新时代的活化传承，不仅是传承中华文明、开创新时代语言文字新局面的迫切要求，也是学术界一直以来探索和实践的方向。

机器学习、深度学习等技术的迅速发展在给甲骨文识别带来新的机遇的同时，也提出了更加多元的需求，越来越多的学者开始关注到甲骨文识别与古籍文字资源的深入挖掘整合与多途径传播。

已有的文献 [6-7] 大多从计算机视觉角度出发对甲骨文识别技术进行综述，缺乏在数字人文视域下对甲骨文识别的前沿热点探讨。鉴于此，笔者采用系统性综述的方法对截至 2022 年上半年的国内外 155 篇针对甲骨文识别研究的文献进行梳理、归纳和分析，将数字人文理念、技术和方法与甲骨文识别技术相结合，旨在揭示数字人文视域下甲骨文识别的研究现状，分析难点与挑战，进而分析发展方向，助力甲骨文识别技术的发展，为甲骨文的活化利用、古籍特定领域数字人文研究提供支撑，促进数字人文研究有效发展，拓宽数字人文边界，同时帮助有关学者挖掘古籍文字的多维价值，促进中华文明的传承发展。

1 甲骨文识别研究现状

1.1 数据来源与研究方法

本研究主要采用文献计量分析法，在多个数据库中通过特定检索式，检索获得多篇相关文献，同时借助 VOSviewer、Excel 等可视化工具从宏观层面把握甲骨文识别技术研究发展现状，既可以在时间上分析相关主题的发展历程，也可以系统地分析数字人文视域下甲骨文识别技术的研究重点与方向。

在 Web of Science、谷歌学术数据库中通过高级检索，运用检索式 TS=(‘oracle bone script’ or ‘Oracle’ or ‘oracle bone’ or ‘oracle bone inscriptions’) AND TS=(‘recognition’ or ‘detection’) 检索英文文献。同时，在中国知网数据库中运用检索式 SU= 甲骨文识别 OR SU= 甲骨文检测 OR SU=(‘甲骨文’+’甲骨文拓片’)*(‘识别’+’检测’) OR KY=(‘甲骨文’+’甲骨文拓片’)*(‘识别’+’检测’) OR (AB=(‘甲骨文’+’甲骨文拓片’)*(‘识别’+’检测’) and KY=(‘识别’+’检测’)) 检索中文文献，筛选截至 2022 年上半年的近几十年来的文献，经过人工筛选，剔除与甲骨文识别技术主题无关的文献，最终获取 103 篇英文文献和 52 篇中文文献。

检索结果所得论文年发文量如图 1 所示。从图 1 可以看出，学界对于甲骨文识别技术的相关研究热度逐渐增加，论文年发文量在近 5 年呈现较快增长，对甲骨文识别进行系统性综述有较高的研究价值。

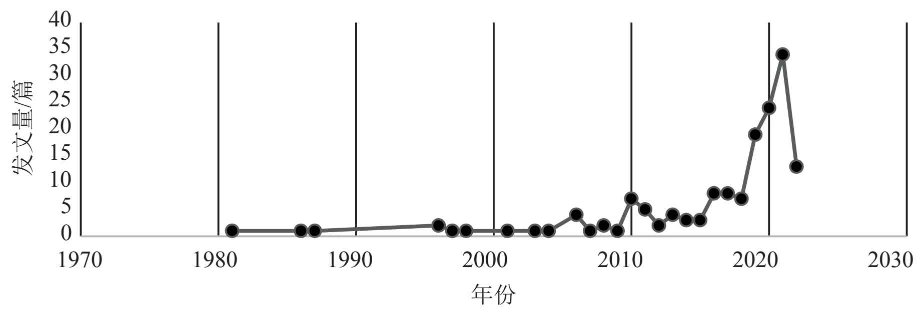


图 1 论文年发文量

知识管理论坛，2023(2): 115-125
DOI: 10.13266/j.issn.2095-5472.2023.010

通过 VOSviewer，对 103 篇英文文献的标题和摘要内容进行聚类分析，从主题与时间两个维度绘制聚类图，可以展现当前甲骨文识别领域的研究热点与研究趋势，如图 2、图 3 所示：

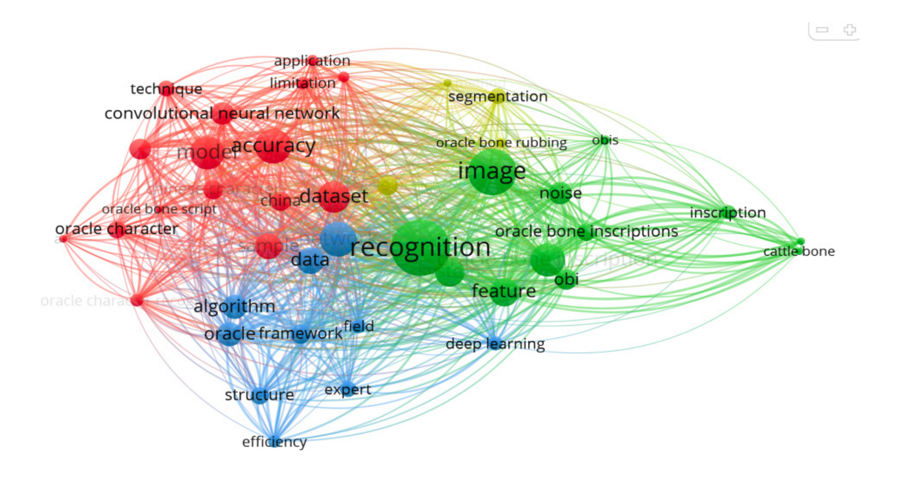


图 2 数字人文视域下甲骨文识别研究论文聚类

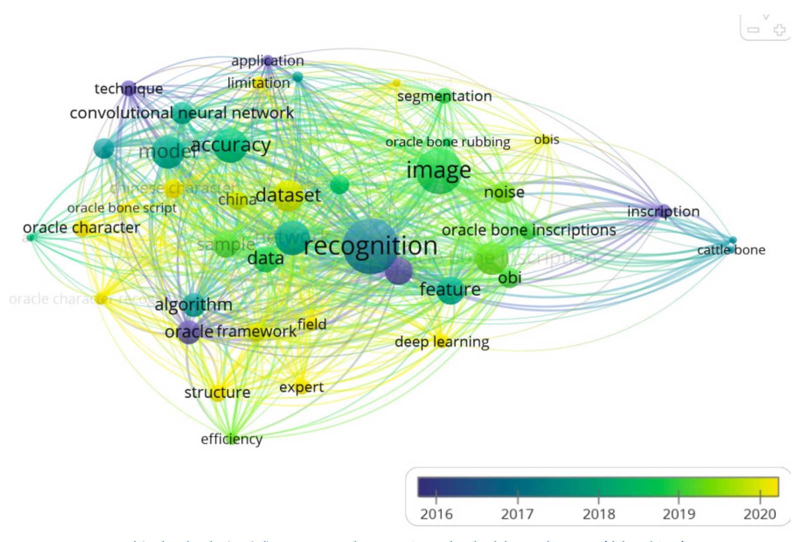


图 3 数字人文视域下甲骨文识别研究文献聚类图（按时间）

从图 2、图 3 可以看出，在研究对象上甲骨文识别主要针对甲骨文拓片图像、甲骨文字形，在识别技术上主要依赖深度学习、构建模型，准确性、速度效率为该领域重要的评价指标。随着技术的发展成熟，深度学习技术的高度应用与数据库、知识库的搭建逐渐成为学界关注的方向。

1.2 甲骨文识别技术研究现状

1.2.1 传统识别技术

传统的甲骨文识别技术主要采用特征提取（feature extract）的方法，基于甲骨文字形结构特征，从字形特征角度或拓扑图形特征角度出发，对其特征进行提取，从而进行分级甲骨文识别。

由于甲骨文字形复杂多变，之前的研究者将其视作线条图，从字形特征出发，通过分析笔划方向、弯曲度、折弯程度等，来提取拓扑特征。复旦大学的周新伦和李锋等 [8-9] 把甲骨文看作是由“线”与“点”构成的无向图，进行图特征提取，将各类图论编码作为字符的拓扑

特征，从而进行分级识别。

但甲骨文字大多是刻在硬骨甲壳上，噪声较大，前者在去噪识别特征方面精确度较低，而基于轮廓线的特征提取、描述可以提高甲骨文识别的精确度，因此后者主要从图形特征或拓扑结构出发进行甲骨文识别。2010 年吕肖庆、李沫楠等 [10] 将基于曲率直方图的傅里叶描述子（Fourier descriptor of curvature histogram, FDCH）作为新的特征，并据此计算出甲骨文字对应的特征向量，从而计算相似度进行甲骨文识别。2016 年顾绍通 [11] 通过分析甲骨文字形的拓扑特征，将甲骨文转化为拓扑图形，对其进行编码，将拓扑编码与拓扑特征库进行配准，从而实现甲骨文识别。

1.2.2 基于机器学习识别技术

由于机器学习算法在计算机视觉领域获得了很大成功，因此除了图论方法中手动编码进行匹配之外，部分甲骨文字形识别工作也引入了机器学习算法。

支持向量机（Support Vector Machine, SVM）[12] 是隶属机器学习的新一代学习方法，广泛应用于图像分类、手写图像识别等领域。与人工神经网络相比，SVM 对测试样本具有更好的泛化能力，对于资源稀缺的甲骨文识别具有优势。X. Shi[13] 利用语料库相关技术处理形成了一个简单的甲骨文数据库，并在此基础上进行数据挖掘，使用 SVM 进行基于部首的分类，以达到知识共享和辅助甲骨文学者考证的目的。刘永革等 [14] 通过块状直方图提取特征，引入经典的 SVM 作为甲骨文识别的模型，使精确度达到了 88%。此外，度量学习在克服甲骨文识别中训练数据的局限性和不平衡性方面也有较多应用。

1.2.3 基于深度学习识别技术

机器学习需要人为机器提供特征进行学习，对应到甲骨文识别中，即需要先定义不同甲骨文类的特征，对专家的依赖性较高而导致信息共享普及率较低，在识别精度、识别效率等方面仍有提升的空间。将深度学习技术运用到甲骨文识别中，能够利用大量甲骨文字形数据集自动提取拓扑特征，即不需要人为定义特征和规则，交由计算机独立识别图像，并根据给定数据图像进行自我迭代训练，从而提高甲骨文识别精度与识别效率。

深度学习技术在甲骨文识别领域的应用主要可以分为两步：第一步是实现甲骨文字体的目标检测及实例分割，第二步是识别被提取的甲骨文字符。关于甲骨文识别方向的研究除了提升甲骨文识别精度以外，还包括低资源甲骨文字符识别、甲骨文变体识别等。近年来，学界对甲骨文拓片图像识别的研究逐渐增多，且识别效果较好，相关研究统计见表 1。

（1）甲骨文检测。Faster R-CNN 是目标检测领域的代表性算法，在此基础上 Z. Liu 等 [27] 优化了 Faster R-CNN 负样本过多的问题，大大提高了检测精度；L. Meng 等 [21] 使用并扩展了单次多箱探测器（Single Shot MultiBox Detector, SSD）[28] 来检测带有摩擦的甲骨文字符，改进了 SSD 在识别较小物体时准确度较低的问题; X. Yue 等 [25] 应用一种基于形态学的分割方法对白川手写甲骨文文档中的字符进行分割，并提出一种神经网络来消除错误分割字符的噪声。

（2）甲骨文识别。基于深度学习的甲骨文识别技术将传统方法的特征提取与多种处理过程合为一体，依靠大量的训练数据和强大的计算性能，在甲骨文识别中逐渐发挥重要作用。刘芳、李华飙等 [16] 基于 Mask R-CNN 进行甲骨文拓片识别，识别准确率提升至 95%；闫升、刘芳等 [17] 进一步改进 Mask R-CNN，实现类别屏蔽与自动识别较正相结合，首次针对拓片图像进行甲骨文字符检测与识别一体化；林小渝等 [18,29] 在深度学习模型的基础上，首次提出从甲骨文单偏旁角度进行甲骨文识别，不仅取得较高的识别率，还能帮助识别甲骨文新字，即零样本学习（zero-shot learning），具有较高的应用意义。

知识管理论坛，2023(2): 115-125
DOI: 10.13266/j.issn.2095-5472.2023.010

表 1 用于深度学习的拓片图像识别数据统计表

作者	数据集	技术方法	结果
陈婷珠、吴少腾等 [15]	《殷墟小屯村中村南甲骨》：515 片甲骨，6 230 张甲骨单字图像	将甲骨文图像转化为编码	训练集：100%
刘芳、李华飙等 [16]	《甲骨文合集》：4 378 张甲骨文单字图像	Mask R-CNN	检测和识别准确率均达到 95%
闫升、刘芳等 [17]	中国国家博物馆馆藏甲骨精品拓片图像以及《甲骨文合集》中部分甲骨拓片图像和入《甲骨文常用字字典》辅助数据集	改进 Mask R-CNN，实现检测与识别一体化	训练集：99.5% 测试集：61.7%
林小渝、陈善雄等 [18]	HCL2000 数据集	甲骨文偏旁：BN-LeNet 网络；甲骨文合体字：OraNet 模型	甲骨文偏旁：96.24%； 甲骨文合体字：98.58%
张颐康、张恒等 [19]	安阳师范学院甲骨文信息处理实验室甲骨文数据集（共 295 466 个样本），选取 241 类拓片甲骨文字样本	跨模态深度度量学习	已知：86.7%； 新类：62.1%
L. Meng、N. Kamitoku 等 [20]	由“上海博物馆甲骨文字”扫描而来	自上向下扩展聚类（Top-Down Extension Clustering, TDE-C) 和依赖矩阵 (Dependency Matrix, DM)	首次使用深度学习识别真实甲骨文字符，准确率达到 92.3%
L. Meng、B. Lyu 等 [21]	一个由真实摩擦图像组成的甲骨文数据集（同类中的第一个数据集）	单侧多箱检测器（Single Shot MultiBox Detector, SSD）	准确率达到 95%
L. Meng、N. Kamitoku 等 [22]	由“上海博物馆甲骨文字”扫描而来	SSD	准确率达到 97%
N. Wang、Q. Sun 等 [23]	公开网络数据集“殷契文渊”	YOLOv4 模型（You Only Look Once version4）	识别准确率达到 75%，召回率达到 90%
B. Du、G. Liu、W. Ge[24]	公开网络数据集“殷契文渊”	双分支自我监督深度学习	
X. Yue、B. Lyu 等 [25]	立命馆大学白川静香东亚文字文化研究所“白川字体”	动态 K-means 聚类、定向梯度直方图（Histogram of oriented gradient, HOG）特征、神经网络	区分噪声和字符的准确率达到 96.5%，字符分类准确率达到 74.91%
C. S. Zhang、R. X. Zong 等 [26]	真实甲骨文数据集 OB-Rejoin、甲骨文注释数据集 OracleBone-8000	甲骨文重联算法、基于深度学习的场景文本检测算法、深度模型匹配算法	碎片匹配前 10% 准确率为 98.39%；甲骨文定位 F 得分为 89.7%；甲骨文识别总体准确率为 80.9%

Z. Guo 等 [30] 提出一种基于 Inception-v3 的用于甲骨文识别神经网络模型，该模型比 AlexNet、VGG-19 更加优越，在特征模糊、遮挡、残缺的情况下仍能取得良好的效果；藤川等 [31] 提出了一种两阶段方法，采用最新的“只看一次”（YOLO）模型和 MobileNet 进行带有摩擦的甲骨文字符识别。这些方法引入了神经网络和深度学习，使模型获得了更好的特征表示能力，因此字符识别的准确性得到显著提高。由于甲骨文拓片图像训练样本较少、图像磨损较大，因此基于拓片载体的甲骨文识别精确度较低。张颐康等 [19] 创新性地提出基于跨模态深度度量学习的甲骨文识别技术，它改编自 J. Guo 等 [32] 提出的基于卷积神经网络（convolutional neural networks, CNN）的甲骨文识别，在 CNN 和深度度量学习的基础上，配有临摹、拓扑甲

骨文字特征编码器，实现跨模态特征空间建模，最终实现甲骨文识别，将精确度从单模态识别的 66.6% 提升至跨模态识别的 88.4%。

也有学者提出新的甲骨文识别思路。F. Gao 等 [33] 提出了一种基于生成对抗网络的图像从甲骨文到现代汉字的图像翻译方法，首次尝试捕获甲骨文字符图像与现代汉字之间的隐形关系；W. Han 等 [34] 将自我监督学习的思想融入到数据增强中，在识别很少拍摄的甲骨文字符时具有较高的性能。

（3）低资源字符识别。标注语料稀缺且分布不平衡，部分甲骨文字符只有一个或几个基础样本，这种带标注训练语料不足条件下的识别任务被称为低资源识别任务，直接使用深度学习方法不能很好地识别低资源字体。因此，J. Li 等 [35] 提出了一种混淆策略，利用混合多数类和少数类的方法来增加样本，并使用三重损失函数来克服分布不平衡的问题。同时为了避免在数据集小、图像质量低的情况下模型数据过度拟合，L. Dazheng 等 [36] 提出了随机多边形覆盖算法的数据增强算法来模拟训练数据集中可能的损伤对象和数据丢失。

严格意义上的甲骨铭文总数为 3 085 个，占甲骨铭文总数的 51.91%[33]，因此识别变体对于甲骨文研究至关重要。J. Gao 等 [37] 提出了一种两阶段方法来区分它们，在第一阶段通过计算机相关方法识别甲骨文变体字符，然后在第二阶段通过结合先验知识的多域方法进一步识别未识别的甲骨文变体字符；G. Liu 等 [38] 提出通过将深度卷积神经网络（deep convolutional neural network, DCNN）与频谱聚类相结合来识别甲骨文的变体。前者用于为甲骨文图像提供准确的描述，后者用于查找每个甲骨文的变体。

2 甲骨文数据处理与存储

利用知识库、人工智能等多方面新兴技术，并辅之人工复校，既可提升古籍文字识别的准确率，也可充当工具库为数字人文研究提供帮助。甲骨文数据库、知识库的构建，不仅为计算机识别甲骨文提供大量矢量字形，扩充数据的多样性，也为甲骨文各项研究提供丰富的检索帮助，便于推动甲骨文古籍数字化研究，愈发成为当前学界关注的重点。

2.1 数据库的构建和标注

目前，有多家学术机构开展了甲骨文数据库与知识库的构建工作。香港汉达文库 [39] 甲骨文库是最早的甲骨文数据库，目前最大的甲骨文数据库是陈年福构建的甲骨文原文释文数据库 [40]。栗青生和吴琴霞等 [41-42] 为了解决对甲骨文异形字编码与输入的问题，通过有向笔段和笔元描述甲骨文字形，并建立甲骨文字形动态描述库，这也有助于甲骨文识别。

随着人工智能等技术的突破，机器学习、深度学习逐渐融入甲骨文字识别等古籍数字化工作中，助力数字人文研究。多位学者 [43] 提出基于人工智能技术训练深度学习模型，并在此基础上建立甲骨文字形数据库，以此帮助甲骨文字检索。S. Huang 等 [44] 构建了一个名为 OBC306 的甲骨文字符大型数据集，并基于标准的深度 CNN 对该数据集进行评估，作为甲骨文识别的基准模型。

在现有的技术环境中，只有经验丰富的甲骨文专家才能对甲骨文进行手动注释，这不仅耗费人力资源，而且效率低下。针对这一问题，S. H. I. Xian-Jin 等 [45] 在甲骨文图像识别模型的基础上，提出一种基于锚点的甲骨文字符级图像自动注释算法。

2.2 领域知识的建立

甲骨文知识库与知识图谱是甲骨文数据库的扩展，是在甲骨文数据库、文字库的基础上，进行条件概率语法现象统计、甲骨文语料分析、句法分析等之后建立的综合知识库，用以进行知识组织与知识服务。建立甲骨文文字库和综合智能知识库，支持逐级排歧校正，有助于准确表达甲骨文含义，助力数字人文研究，也为甲骨文信息处理提供创新性的研究思路 [46-47]。

J. Xiong 等 [46] 针对甲骨文研究学习难度大、学习周期长、知识点广但知识连接弱、共享度

知识管理论坛，2023(2): 115-125
DOI: 10.13266/j.issn.2095-5472.2023.010

低等问题，提出一种构建多模态知识图谱的解决方案。甲骨文多模态知识图谱可以为多源异构数据提供统一的语义空间。通过多模态融合和信息互补，可以解决信息处理中单一模态的缺陷。这个多模态知识图谱可以更好地组织和管理基础数据，为甲骨文信息处理研究服务。

安阳师范学院是国内唯一的甲骨文理工科研究基地，与社会科学院甲骨学殷商史研究中心共同建设“三库一平台”，即甲骨文字库、著录库、文献库和甲骨文知识服务平台，标志着甲骨学研究由“数字化”进入“智能化”时代 [48]。其中大数据平台构建了基于人工手写甲骨文字符数据库 hwobc，它包含 83 245 个字符级样本，3 881 个字符类别，并采用传统深度学习分类网络进行学习分类。一方面深度学习打破馆藏资源的界限，公开扩大数据集资源，从而形成丰富的测试集，提升深度学习的性能；另一方面实现文史研究与智能技术的深度融合，促进甲骨文研究工作的发展。

在领域知识的建立中，知识本体可以以知识元的形式对智能技术提取出的数据进行有效关联，构建出语义网络，提高对数据资源的整合利用，同时语义网络也可利用其推理、计算能力，帮助研究者考释未破译的甲骨文字 [49]。例如，Q. Jiao 等 [50] 构建语义网络，进行具有相似语义的甲骨文字符的模块结构检测。

3 现有不足

3.1 数据特征

甲骨文的构成方式主要为 4 种，分别是象形、形声、会意和指事。其中，象形字占据了较大比例，一些形声字、会意字也是在象形字的基础上发展而来 [11]。因此，甲骨文字具有较强的图画性。现阶段，相关领域的大多数学者倾向于将甲骨文归类至图形体文字而非笔画体文字。他们认为，甲骨文不仅不存在现代汉字中所谓的笔画概念，在笔画多少、正反向背等方面也没有统一要求。甲骨文偏旁部首的排列既不是横排也不是竖排，在字形结构上有着一定的随意性。同时，由于甲骨文笔端尖细、难以区分笔画，专家在识别甲骨文时只能将其作为一个整体输入。这些特点在学者采用现代化技术对其识别时造成了较大的困扰。

由于兽骨、龟甲上可供镌刻、书写的位置有限，以现代标准来衡量，甲骨文的排版是参差错落、疏密不均、大小不一的，部分甲骨文字为了能够更加准确地表示相对复杂的实物，一个字通常会占据多个字的位置 [10]。因此，在对甲骨文进行识别的过程中少有版式信息可以借助。

类比于现代汉字的书写系统，不同的人对于同一个甲骨文字也有着多种不同的刻写方法。例如，一些会意字只需要指定偏旁结合就能够表示某种含义，而不要求其位置固定 [10]。不同的刻写方法造成了不同形体的甲骨文的存在，不同形体的甲骨文之间差别很大 [51]。字体变体和相似字符之间的混淆使得甲骨文的识别具有一定的难度。此外，甲骨文字频存在两端集中现象，即少数高频字占总字量的高比重，和在总字量中占极低比重的低频字占单字总数的极高比重 [52]。低频字高度集中的现象表现出甲骨文作为一个文字系统的不成熟性。除此之外，还有大量的甲骨文属于未考释字 [15]，这些特征都为甲骨文的识别增加了难度。

部分甲骨拓片受到年代久远、保存条件恶劣等因素的影响，表面遭受不同程度的残蚀与破损。考古学家在获取拓片甲骨文字图像的过程中也会对原始甲骨拓片产生一定的破坏，如去除拓片上的残泐痕和其他文字的痕迹等 [43]，这些操作可能会导致甲骨文字缺笔变形。因此，大部分拓片甲骨文字图像都具有图像残缺、噪声严重等缺点。

3.2 识别技术

甲骨文识别技术目前尚处于起步阶段，现有的甲骨文识别技术不仅存在无法完全提取甲骨文字的特征、无法完全符合甲骨文字的实际情况等问题，其本身的复杂性也使现有算法在使用范围等方面受到一系列的限制。换言之，

目前甲骨文识别技术的性能还不太能够达到完全实用化水平，未来有待进一步发展与完善。

以卷积神经网络为核心的深度学习技术在大数据环境下能够取得较为理想的甲骨文识别效果，但该种技术并未充分利用甲骨文的自身特征，无法为神经网络提供大量的特征提取样本，在其他条件下的识别效果不尽人意。

文字识别领域性能优异的深度学习方法对大量样本训练有着较高程度的依赖。因客观条件的限制，获取拓片甲骨文字具有较大的难度，这导致深度学习方法缺乏训练样本，深度学习算法在训练集样本足够大的情况下才能充分发挥其性能，而甲骨文样本数量少，历史跨度大，字形演变丰富，数据集不充分 [16]。因此，该种方法对真实的拓片资源很难取得较高的识别精度。

在目前出土的甲骨拓片中，大部分甲骨文的字形无法得到准确辨识，其读音和意义仍待进一步考究，这使得甲骨文编码输入的方法存在规则繁重、重码多和识别效率低的缺陷 [51]。以史小松为代表的“甲骨文字结构派”学者采用语料库和支持向量机的理论并建立了甲骨文字形库和语料库，但该方法不仅在识别图画特征明显、结构不清晰的甲骨文字时存在困难，还伴有识别效率低的问题。

4 甲骨文识别的未来工作

4.1 数据的扩展

安阳师范学院和中国社会科学院甲骨学殷商史研究中心合作建设的甲骨文大数据及资料检索分析平台“殷契文渊” [53] 中涉及国内外多家机构的原始甲骨文拓片图像，在一定程度上实现甲骨文拓片资源共享，帮助甲骨学资源由“独享”到“共享”，提供更多的原始拓片数据集，提高数据量与覆盖度。而要进一步推动甲骨文识别研究，需要进一步拓宽这种资源共享的渠道，该项工作任重而道远。

数据集中样本数量的缺少会导致识别精度较低，同时由于甲骨文原始资源大多存在图像残缺、背景噪声严重的问题，因此当一个甲骨文字符写入时可以考虑从字符的角度或厚度出发，通过顺时针（clockwise rotate）或逆时针旋转（counterclockwise rotate）、字符加深（dilate）或腐蚀化（erode）、压缩（compress）或拉伸（stretch）等操作，经过多次转化生成新的图像，由此扩展数据集。

在将甲骨文数据信息转移到电脑与网络的过程中，无论是编码类输入法还是无编码类输入法均需要足够的甲骨文专业知识，且对于未破译的甲骨文字需要逐个检索甲骨文字形描述库，这无疑造成甲骨学研究的巨大障碍。因此，应当提升甲骨文输入法技术，实现零学习成本与输入效率的双赢，使数字人文中的古籍数字化研究更便利，也更有利于甲骨文的研究与发展。

4.2 技术的优化

甲骨文虽是较成熟的文字系统，但仍处于汉字早期阶段，异体字众多、低频字高度集中，大量实验存在检测正确但识别错误的情况，易出现分类过度的问题，仍需要专家复审，对专家的依赖度较高。甲骨文识别研究可从数据增强、模型结构调整、优化实现 3 个方向提高识别精度。当前数据增广策略的研究对象基本为拓片图像，可进一步利用甲骨文单字进行研究。因此在日后的研究中，该领域研究者应考虑数据的噪声、图像残缺和算法的泛化能力弱等问题，加快技术开发，提高针对原始甲骨文拓片资源的识别效率。针对卷积神经网络本身，网络深度过多会导致梯度消失或爆炸的问题，从而导致网络性能下降，同时网络深度也不容易训练，因此不需要选择更深入的神经网络，而是采用最合适的优化方法。

数字人文是将信息技术、数字技术融入传统的人文社科研究，数字人文研究者同时具备工具、数据与人文社科理论，应当逐步做到文本分析、文化分析 [54-55]，超越简单的文字阐释。因此，在甲骨文知识库、知识图谱的构建中，应更多考虑提取拓片全文，抽取更多实体与关

知识管理论坛，2023(2): 115-125
DOI: 10.13266/j.issn.2095-5472.2023.010

系，而非仅仅依靠元数据信息抽取，从而建立更完善的甲骨文知识关联网络，拓宽甲骨文考释研究，从“数据化”“数字化”发展为“智能化”，加强甲骨文资源数据库与智能深度识别甲骨文字信息应用平台的建设。

5 结语

本研究对国内外甲骨文识别研究现状和发展动态进行了述评，并在数字人文视阈下探讨相关热点。纵观当前研究成果，随着技术的发展应用，甲骨文识别技术从传统的特征提取到基于深度学习的各类技术，发展迅速且前景广阔。展望未来，数字人文视阈下的甲骨文识别的发展具有较高的研究意义。提升甲骨文识别技术，提高甲骨文分类率，构建甲骨文知识库和知识图谱，建立领域知识，这些都将成为甲骨文识别研究的重要内容和重要命题，研究结果也将为新时代甲骨文的探索和实践提供重要的理论指导和工具。

参考文献：

[1] 沃尔什 , 科布 , 弗雷默里 , 等 .iSchool 中的数字人文 [J]. 陈怡 , 译 . 数字人文研究 , 2021, 1(3): 93-112.

[2] 邓君 , 王阮 . 数字人文视域下口述历史档案资源知识发现模型构建 [J]. 档案学研究 , 2022(1): 110-116.

[3] 李巧明 , 王晓光 . 跨学科视角下数字人文研究中心的组织与运作 [J]. 数字图书馆论坛 , 2013(3): 26-31.

[4] 陈力 . 数字人文视域下的古籍数字化与古典知识库建设问题 [J]. 中国图书馆学报 , 2022, 48(2): 36-46.

[5] 刘乾先 , 董莲池 , 张玉春 , 等 . 中华文明实录 [M]. 哈尔滨 : 黑龙江人民出版社 , 2002.

[6] 卢芯怡 . 新时期甲骨文应用研究述评 [J]. 汉字文化 , 2020(21): 73-78.

[7] 刘国英 . 基于深度学习的甲骨文字检测与识别 [J]. 殷都学刊 , 2020, 41(3): 54-59.

[8] 李锋 , 周新伦 . 甲骨文自动识别的图论方法 [J]. 电子科学学刊 , 1996(S1): 41-47.

[9] 周新伦 , 李锋 , 华星城 , 等 . 甲骨文计算机识别方法研究 [J]. 复旦学报 (自然科学版), 1996(5): 481-486.

[10] 吕肖庆 , 李沫楠 , 蔡凯伟 , 等 . 一种基于图形识别的甲骨文分类方法 [J]. 北京信息科技大学学报 (自然科学版), 2010, 25(S2): 92-96.

[11] 顾绍通 . 基于拓扑配准的甲骨文字形识别方法 [J]. 计算机与数字工程 , 2016, 44(10): 2001-2006.

[12] CRISTIANINI N, TAYLOR J S. 支持向量机导论 [M]. 李国正 , 王猛 , 曾华军 , 译 . 北京 : 电子工业出版社 , 2004.

[13] SHI X. Research on oracle word structure analysis based on support vector machine[D]. Shanghai: East China Normal University, 2010.

[14] LIU Y, LIU G. Oracle-bone inscription recognition based on svm[J]. Journal of Anyang Normal University, 2017, 2: 54-56.

[15] 陈婷珠 , 吴少腾 , 吴江 , 等 . 基于编码的甲骨文识别技术研究 [J]. 中国文字研究 , 2019(1): 1-12.

[16] 刘芳 , 李华飙 , 马晋 , 等 . 基于 Mask R-CNN 的甲骨文拓片的自动检测与识别研究 [J]. 数据分析与知识发现 , 2021, 5(12): 88-97.

[17] 闫升 , 刘芳 , 孙岱萌 , 等 . 博物馆基于人工智能的甲骨文知识普及与活化传承 [J]. 中国博物馆 , 2021(3): 110-116, 144.

[18] 林小渝 , 陈善雄 , 高未泽 , 等 . 基于深度学习的甲骨文偏旁与合体字的识别研究 [J]. 南京师大学报 (自然科学版), 2021, 44(2): 104-116.

[19] 张颐康 , 张恒 , 刘永革 , 等 . 基于跨模态深度度量学习的甲骨文字识别 [J]. 自动化学报 , 2021, 47(4): 791-800.

[20] MENG L, KAMITOKU N, YAMAZAKI K. Recognition of oracle bone inscriptions using deep learning based on data augmentation[C]//2018 metrology for archaeology and cultural heritage (MetroArchaeo). Piscataway: IEEE, 2018: 33-38.

[21] MENG L, LYU B, ZHANG Z, et al. Oracle bone inscription detector based on SSD[C]//International conference on image analysis and processing. Berlin: Springer, 2019: 126-136.

[22] MENG L, KAMITOKU N, KONG X, et al. Deep learning based ancient literature recognition and preservation[C]//2019 58th annual conference of the Society of Instrument and Control Engineers of Japan (SICE). Piscataway: IEEE, 2019: 473-476.

[23] WANG N, SUN Q, JIAO Q, et al. Oracle bone inscriptions detection in rubbings based on deep learning[C]//2020 IEEE 9th joint international information technology and artificial intelligence conference (ITAIC). Piscataway: IEEE, 2020: 1671-1674.

[24] DU B, LIU G, GE W. Deep self-supervised learning for Oracle bone inscriptions features representation[C]//2021

IEEE 4th international conference on information systems and computer aided education (ICISCAE). Piscataway: IEEE, 2021: 7-11.
[25] YUE X, LYU B, LI H, et al. Deep learning and image processing combined organization of Shirakawa's hand-notated documents on OBI research[C]//2021 IEEE international conference on networking, sensing and control (ICNSC). Piscataway: IEEE, 2021: 1-6.
[26] ZHANG C, ZONG R, CAO S, et al. AI-powered oracle bone inscriptions recognition and fragments rejoining[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, 2021: 5309-5311.
[27] LIU Z, WANG X, YANG C, et al. Oracle character detection based on improved faster R-CNN[C]//2021 international conference on intelligent transportation, big data & smart city (ICITBS). Piscataway: IEEE, 2021: 697-700.
[28] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: Proceedings of 14th European Conference. Burlin: Springer, 2016: 21-37.
[29] 林小渝 . 基于深度学习的甲骨文偏旁与合体字识别的研究与实现 [D]. 重庆 : 西南大学 , 2021.
[30] GUO Z, ZHOU Z, LIU B, et al. An improved neural network model based on inception-v3 for Oracle bone inscription character recognition[J/OL]. Scientific programming, 2022[2023-01-27]. https: //doi.org/10.1155/2022/7490363.
[31] FUJIKAWA Y, LI H, YUE X, et al. Recognition of oracle bone inscriptions by using two deep learning models[J/OL]. International journal of dental hygiene, 2022[2023-01-27]. https: //doi.org/10.1007/s42803-022-00044-9.
[32] GUO J, WANG C H, ROMAN-RANGEL E, et al. Building hierarchical representations for oracle character and sketch recognition[J]. IEEE transactions on image processing, 2016, 25(1): 104−118.
[33] GAO F, ZHANG J, LIU Y, et al. Image translation for oracle bone character interpretation[J]. Symmetry, 2022, 14(4): 743.
[34] HAN W, REN X, LIN H, et al. Self-supervised learning of orc-bert augmentator for recognizing few-shot oracle characters[C]//Proceedings of the Asian conference on computer vision. Kyoto: Revised Selected Papers, 2020: 652-668.
[35] LI J, WANG Q F, ZHANG R, et al. Mix-up augmentation for oracle character recognition with imbalanced data distribution[C]//Document analysis and recognition–ICDAR 2021: 16th international conference. Berlin: Springer International Publishing, 2021: 237-251.
[36] DAZHENG L. Random polygon cover for Oracle bone character recognition[C]//2021 5th international conference on computer science and artificial intelligence. New York: Association for Computing Machinery, 2021: 138-142.
[37] GAO J, LIANG X. Distinguishing oracle variants based on the isomorphism and symmetry invariances of oracle-bone inscriptions[J]. IEEE access, 2020, 8: 152258-152275.
[38] LIU G, GE W, DU B. Recognition of OBIC's variants by using deep neural networks and spectral clustering[C]//2021 IEEE 4th international conference on information systems and computer aided education (ICISCAE). Piscataway: IEEE, 2021: 39-42.
[39] 杨琳 . 数字化古典文献综述 [J]. 中国史研究动态 , 2004(4): 20-27.
[40] 门艺 . 由甲骨学工具书的编纂到甲骨文数据库的建设 [J]. 漯河职业技术学院学报 , 2019, 18(5): 1-7.
[41] 栗青生 , 吴琴霞 , 王蕾 . 基于甲骨文字形动态描述库的甲骨文输入方法 [J]. 中文信息学报 , 2012, 26(4): 28-33.
[42] 栗青生 , 吴琴霞 , 杨玉星 . 甲骨文字形动态描述库及其字形生成技术研究 [J]. 北京大学学报 (自然科学版), 2013, 49(1): 61-67.
[43] 门艺 , 张重生 . 基于人工智能的甲骨文识别技术与字形数据库构建 [J]. 中国文字研究 , 2021(1): 9-16.
[44] HUANG S, WANG H, LIU Y, et al. Obc306: a large-scale oracle bone character recognition dataset[C]//2019 international conference on document analysis and recognition (ICDAR). Piscataway: IEEE, 2019: 681-688.
[45] XIAN-JIN S H I, SHUANG C A O, CHONG-SHENG Z, et al. Research on automatic annotation algorithm for character-level oracle-bone images based on anchor points[J]. Acta electonica SINICA, 2021, 49(10): 2020-2031.
[46] XIONG J, LIU G, LIU Y, et al. Oracle bone inscriptions information processing based on multi-modal knowledge graph[J]. Computers & electrical engineering, 2021, 92: 107173.
[47] 江铭虎 , 邓北星 , 廖盼盼 , 等 . 甲骨文字库与智能知识库的建立 [J]. 计算机工程与应用 , 2004(4): 45-47, 60.

知识管理论坛，2023(2): 115-125
DOI: 10.13266/j.issn.2095-5472.2023.010

[48] 甲骨文信息处理重点实验室 [EB/OL]. [2021-04-09]. http: //jgwsys. aynu.edu.cn/index.htm.
[49] 熊晶 , 韩胜伟 . 甲骨文研究中跨模态知识图谱的重要性刍议 [J]. 殷都学刊 , 2020, 41(3): 60-64, 97.
[50] JIAO Q, JIN Y, LIU Y, et al. Module structure detection of oracle characters with similar semantics[J]. Alexandria engineering journal, 2021, 60(5): 4819-4828.
[51] 顾绍通 . 基于分形几何的甲骨文字形识别方法 [J]. 中文信息学报 , 2018, 32(10): 138-142.
[52] 刘志基 . 简论甲骨文字频的两端集中现象 [J]. 语言研究 , 2010, 30(4): 114-122.
[53] 李邦 , 刘永革 . 文献数字化技术在甲骨文数据库建设中的应用与展望 [J]. 殷都学刊 , 2020, 41(3): 47-53.
[54] 赵薇 . 作为计算批评的数字人文 [J]. 中国文学批评 , 2022(2): 157-166, 192.
[55] LIU A. Where is cultural criticism in the digital humanities?[M]. GOLD M K. Debates in the digital humanities. Minneapolis: University of Minnesota Press, 2012: 495-501.

作者贡献说明：
刘　洋：确定选题，提出研究思路，修改论文；
陆　逸：分析和处理数据，撰写论文；
魏钰驰：分析和处理数据，撰写论文；
孙智莹：分析和处理数据，撰写论文；
朱立芳：修改论文。

Research Status and Prospect of Oracle Bone Inscription Recognition Technology

Liu Yang[1,2] Lu Yi[1] Wei Yuchi[1] Sun Zhiying[1] Zhu Lifang[3]
[1]School of Information Management, Wuhan University, Wuhan 430072
[2]Big Data Research Institute, Wuhan University, Wuhan 430072
[3]School of Humanities and Communication, Guangdong University of Finance and Economics, Guangzhou 510320

Abstract: [Purpose/Significance] Digital humanities research is a prominent research hotspot in the current academic circle. This study systematically reviewed the frontier research on oracle bone inscription recognition from the perspective of digital humanities, which provided reference for follow-up research, promoting the effective development of digital humanities research and the recognition and utilization of characters in ancient books. **[Method/Process]** The literature was retrieved from multiple academic platforms such as WOS and CNKI using the method of bibliometric analysis, and a total of 103 English literature and 52 Chinese literature were screened for review. **[Result/Conclusion]** Interpreting the research status of oracle bone inscription recognition from three levels: traditional recognition technology, machine learning and deep learning, which analyzed the research development process, and discussed the future development trend. This paper mainly conducted a systematic review of oracle bone inscription recognition research from the perspective of digital humanities, which analyzed existing research technologies and research directions, but did not elaborate on the recognition algorithm mechanism in depth. Oracle recognition technology has gradually changed from traditional feature extraction to deep learning-based recognition technology. Although the recognition accuracy has been improved, there are still shortcomings such as serious overfitting and low recognition efficiency. Meanwhile, the construction of oracle knowledge base and knowledge graph, and the establishment of domain knowledge have good development potential in this field.

Keywords: digital humanities　oracle bone recognition　research progress　review